DOCUMENT RESUME

ED 154 342 CS 004 062

AUTHOR Hopkins, Carol J.; Moe, Alden J.
TITLE The Computer-Assisted Identification of Common Word Strings from the Text of Children's Books.
PUB DATE May 78
NOTE 12p.; Paper presented at the Annual Meeting of the International Reading Association (23rd, Houston, Texas, May 1-5, 1978) For related document see CS004061

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Beginning Reading; *Childrens Books; *Computers; *Phrase Structure; Primary Education; Sight Vocabulary; Structural Analysis; *Syntax; *Word Frequency; *Word Lists; Word Recognition
IDENTIFIERS Trade Books; *Word Strings

ABSTRACT

The complete texts of 250 trade bcoks for children in the primary grades were analyzed by computer in order to identify recurring two- and three-word strings. Of the 202,763-word sample that resulted, 89 two-word strings occurred 100 times cr more, and only two three-word strings occurred more than 100 times. These frequencies represent, respectively, 10.5% and 0.15% of the total number of strings in the sample. The investigatcrs propose that these results are directly applicable to classroom reading instruction, and suggest that beginning readers can be taught such word strings in much the same manner as common sight words are presently taught in initial reading instruction. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED154342

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Carol J. Hopkins
205 Education Building
Purdue University
West Lafayette, IN 47907

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Carol J. Hopkins

Alden J. Moe

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

THE COMPUTER-ASSISTED IDENTIFICATION OF COMMON WORD STRINGS FROM THE TEXT OF CHILDREN'S BOOKS

Carol J. Hopkins and Alden J. Moe

Purdue University

A paper presented at the twenty-third annual convention of the International Reading Association, Houston, Texas, May 4, 1978

The concept of phrases or common word strings as they are referred to in this paper, is not new. As early as 1908, Huey noted that, "Unitary recognition of phrases is very common in reading, for mentally the words do not stand apart" (p. 115). In a related statement, he claimed that, "The reader's acquirement of ease and power in reading comes through increasing ability to read in larger units" (p. 116). The need to teach readers to attend to phrases has recently been cited by Harris and Sipay (1971, 1975), Dechant (1970), Dallmann, Rouch, Char and DeBoer (1978).

The purpose of this investigation was to identify common two-word and three-word strings in the text of children's trade books.

CS 004062

Procedures

As stated in the paper by Moe and Hopkins (1978), the investigators believed that in order to substantiate claims that they had identified certain phrases which were used with any degree of regularity in printed materials, a minimum of 50,000 running words needed to be analyzed. Samples of hundreds of thousands of words were thought to be the most desirable. To obtain a language corpus of acceptable size, the investigators selected 250 widely read, highly recommended picture books for children published between the years 1928 and 1977.

The complete text of each of the trade books frequently read by primary-grade children was keypunched, resulting in a sample of 202,763 words. The text sample was then analyzed via the Moe and Hopkins (1978) computer program designed to identify recurring word strings. The program is designed to identify all consecutive two-word and three-word strings occurring in the corpus. Information regarding the total number of two-word and three-word strings encountered and the number of different strings represented is provided on the printout. The frequency with which the strings occur is tabulated beside the listing of each individual word group.

Findings

Within the total corpus there were 170,249 two-word strings. Of this number, 85,511 different strings were identified. However,

only 89 of these different two-word strings occurred 100 times or more in the corpus. The strings, their frequency of occurrence, and the percentage of the total corpus they represent are presented in Table 1.

- -

Insert Table 1 about here

- -

There were 141,018 three-word strings encountered in the corpus, 120,024 of which were different, meaning that only 15% of them occurred more than one time in the text. Only two of the three-word strings occurred more than 100 times. Table 2 contains a listing of the ten most common three-word strings and their frequency of occurrence.

- -

Insert Table 2 about here

- -

If one considers the cumulative frequency of all two-word strings occurring more than 100 times, it can be seen that these 89 strings account for 10.5% of all strings occurring in text. For three-word strings, the two strings that occur more than 100 times account for .15% of the total number of strings in the corpus.

Discussion

There are several observations to be made about the word strings appearing in Tables 1 and 2. First, while the cumulative frequency of the two-word phrases is not exceedingly high, it is of interest to look at the individual words which make up the two-word

phrases occurring 100 times or more. It is possible that there could have been 178 different words represented in these 89 phrases. However, because of the repeated use of certain words, only 68 different words were used and the majority of these words (75%) are within the 100 most frequently occurring words on the Carroll, Davies, Richman (1971) word list. Table 3 contains a listing of these words and their frequency of occurrence.

Insert Table 3 about here

A second observation to be made is that these common word strings are derived from the text of trade books commonly used with primary grade children, books which typically are written without the vocabulary control found in textbooks, particularly basal readers, written for primary-grade students. The investigators suspect that had the corpus been based on text materials used in the primary grades, the common strings would have occurred with greater frequency.

The third observation deals with a comparison of the common strings identified in the present study with the common strings or phrases occurring on the Dolch Sight Phrase Cards (1948). Dolch's phrase cards represent combinations of the 95 commonest nouns and the 220 words on the Dolch basic sight vocabulary list. Of the 144 phrase cards, none of the 76 three-word phrases appeared within the ten most common three-word strings in the present study. It should be pointed out, however, that if one examines the first two words in Dolch's three-word phrases, 35 of these also appear

in the two-word strings occurring in the present investigation. Six of the 68 two-word phrases included in the Dolch cards appear in the list of two-word strings occurring more than 100 times in the present investigation.

The investigators believe that the results of this study are directly applicable to classroom reading instruction. Since a number of common word strings can be identified in text, and were identified in this study, beginning readers can be taught such word strings in much the same manner as common sight words are presently taught in initial reading instruction.

TABLE 1

MOST COMMON TWO-WORD STRINGS

String	Frequency	% of Total Strings Encountered	Cumulative Frequency
in the	1153	.68	0.68
of the	862	.51	1.19
to the	714	.42	1.61
and the	653	.38	1.99
on the	602	.35	2.34
said the	406	.24	2.58
he was	376	.22	2.80
it was	369	.22	3.02
and he	317	.19	3.21
at the	315	.19	3.40
was a	297	.17	3.57
into the	285	.17	3.74
all the	279	.16	3.90
he said	261	.15	4.05
in a	258	.15	4.20
out of	254	.15	4.35
a little	240	.14	4.49
the little	234	.14	4.63
there was	224	.13	4.76
for the	222	.13	4.89
I am	216	.13	5.02
to be	210	.12	5.14
he had	206	.12	5.26
from the	197	.11	5.37
for a	195	.11	5.48
and a	182	.11	5.59
with a	178	.10	5.69
and I	170	.10	5.79
on his	169	.10	5.89
of his	163	.10	5.99
his mother	163	.10	6.09
I have	159	.09	6.18
with the	158	.09	6.27
it is	158	.09	6.36
to see	158	.09	6.45
I will	154	.09	6.54
went to	154	.09	6.63
began to	153	.09	6.72
when he	153	.09	6.81
but the	149	.09	6.90

TABLE 1 - Continued

String	Frequency	% of Total Strings Encountered	Cumulative Frequency
and she	149	.09	6.99
she said	148	.09	7.08
a big	148	.09	7.17
in his	148	.09	7.26
she was	145	.08	7.34
to get	144	.08	7.42
had a	144	.08	7.50
and they	144	.08	7.58
have a	141	.08	7.66
the water	139	.08	7.74
you are	137	.08	7.82
through the	137	.08	7.90
the sun	137	.08	7.98
the other	137	.08	8.06
to go	135	.08	8.14
like a	134	.08	8.22
going to	134	.08	8.30
they were	134	.08	8.38
the moon	134	.08	8.46
up and	132	.08	8.54
on a	129	.08	8.62
down the	128	.08	8.70
the way	128	.08	8.78
he could	126	.07	8.85
and then	126	.07	8.92
came to	123	.07	8.99
over the	123	.07	9.06
when the	121	.07	9.13
as he	121	.07	9.20
the man	119	.07	9.27
then he	119	.07	9.34
the tree	117	.07	9.41
one day	117	.07	9.48
of a	116	.07	9.55
is a	115	.07	9.62
and his	115	.07	9.69
with his	113	.07	9.76
up the	113	.07	9.83
he went	113	.07	9.90
the big	105	.06	9.96
are you	104	.06	10.02
to his	103	.06	10.08

TABLE 1 - continued

String	Frequency	% of Total Strings Encountered	Cumulative Frequency
back to	102	.06	10.20
but he	102	.06	10.20
by the	101	.06	10.26
the door	100	.06	10.32
did not	100	.06	10.38
the sky	100	.06	10.44
do you	100	.06	10.50

TABLE 2

MOST COMMON THREE-WORD STRINGS

String	Frequency	% of Total Strings Encountered	Cumulative Frequency
out of the	113	.08	.08
there was a	101	.07	.15
went to the	59	.04	.19
it was a	58	.04	.23
back to the	47	.03	.26
up and down	40	.03	.29
go to the	39	.03	.32
the end of	37	.03	.35
came to the	36	.02	.37
said the man	36	.02	.39

TABLE 3

FREQUENCY OF WORDS FOUND IN TWO-WORD STRINGS

Word	Frequency (as found in strings occurring 100 times or more)	Word	Frequency (as found in strings occurring 100 times or more)
the	29	am*	1
a	13	be	1
he	12	from	1
to	11	mother*	1
and	9	see	1
his	7	will	1
was	5	began*	1
of	4	get*	1
I	4	water	1
you	3	through*	1
with	3	sun*	1
said	3	other	1
on	3	go*	1
in	3	like	1
it	2	going*	1
little	2	were	1
for	2	moon*	1
had	2	down	1
have	2	way	1
is	2	could	1
went*	2	came*	1
when	2	over	1
but	2	as	1
she	2	man*	1
big*	2	tree*	1
they	2	one	1
are	2	day*	1
up	2	back*	1
then	2	by	1
at	1	door*	1
into	1	did	1
all	1	not	1
out	1	sky*	1
there	1	do	1

*Denotes word not within the 100 most frequently occurring words on the Carroll, Davies, Richman (1971) word list

REFERENCES

Carroll, J. B., Davies, P. and Richman, B. American Heritage Word Frequency Book. Boston: Houghton Mifflin, 1971.

Dallmann, M., Rouch, R. L., Char, L. C. and DeBoer, J. J. The Teaching of Reading, fifth edition, New York: Holt, Rinehart and Winston, 1978.

Dechant, E. V. Improving the Teaching of Reading, second edition. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Dolch, E. W. Sight Phrase Cards. Champaign, Illinois: Garrard Publishing Co., 1948.

Harris, A. J. and Sipay, E. R. Effective Teaching of Reading, second edition. New York: David McKay, 1971.

Harris, A. J. and Sipay, E. R. How to Increase Reading Ability, sixth edition. New York: David McKay, 1975.

Huey, E. B. The Psychology and Pedagogy of Reading. New York: MacMillan, 1908.

Moe, A. J. and Hopkins, C. J. "Parsing Word Strings from Text With a Computer: Implications for Reading Instruction." Paper presented at the meeting of the International Reading Association, Houston, Texas, 1978.